

## GENOME SEQUENCE ANALYSIS OF *SOLANUM LYCOPERSICUM* BY APPLYING SEQUENCE ALIGNMENT METHOD TO DETERMINE THE STATISTICAL SIGNIFICANCE OF AN ALIGNMENT

UMA KUMARI<sup>1</sup> & ASHOK KUMAR CHOUDHARY<sup>2</sup>

<sup>1</sup>Department of Biotechnology, Jharkhand Rai University, Ranchi, Jharkhand, India

<sup>2</sup>Department of Botany, Ranchi University, Ranchi, Jharkhand, India

### ABSTRACT

Sequencing the genome of the crop *Solanum lycopersicum* will also help to identify beneficial genes in other plant relative of the tomato such as potato, pepper. All of these crops are members of the solanaceae or nightshade family, one of the World most important vegetable plants families in term of both economic value and production volume. Developing better tomatoes will also contribute to the quest for global food security. As well as using this new genome information to develop a wide variety of beneficial traits, the(TGRD)the tomato genomic resources database is an online and interactive relational database developed using open sources software. The user-friendly interface for TRGD has been developed using java script and HTML to query and retrieve the data based on userneeds. In sequence alignment is a way of arranging the sequence of DNA, RNA, or protein to identify the functional structural or evolutionary relationship between the sequence. If two sequence is an alignment share a common ancestors, mismatches can be interpreted as a point mutation. Fasta format is a text based format for representing either nucleotide sequence or peptide sequence, in which nucleotide or amino acid are represented using single letter code. Sequence homology is a general term that indicates evolutionary relatedness among sequence. NCBI that provide a common data extraction platform for sequence analysis. sequence similarity is a substitution with similar chemical properties. The clutalw colored alignment also have the colour option in the output results. The colouring residue takes place according to the following physiochemical criteria(Red, blue, green, magenta, and grey colours).In addition to maintaining the gene bank nucleic acid, sequence database, Ncbi provide data retrieval system and computational resources for the analysis of gene bank data and variety of other biological data made available through Ncbi.

**KEYWORDS:** Bioinformatics, Genome Database, TGRD(Tomato Genomic Resources Database), Sequence Analysis, Data Compiled, Sequence Alignment, *Solanum lycopersicum*

Received: Apr 27, 2016; Accepted: May 05, 2016; Published: May 07, 2016; Paper Id.: IJBTRJUN20162

### INTRODUCTION

The aim of tomato genome sequencing is to reveal and explore the genetic variation availability in tomato. Tomato has been selected as a target crop because it is economically one of the most important crop species. The programme can run online from the EBI web server. The sources code executables for window, linux are available from EBI. The clustal series of programe are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequence and for preparing the phylogenetic trees .Taylor Willie, Higgins Des 2000, Bioinformatics. New features include NEXUS and FASTA format output, printing range numbers and fater tree calculations. clustalw originally developed to run on local computers; numerous web server have been setup,

notably at the EBI(European bioinformatics institute). Tomato have been used extensively for genetic studies because of several reason such as its diploid genome, short generation time, efficient transformation technology. The data can be submitted and accessed via the world wide web( Mount .David 2004),. The tomato genome resources database is a interactive relational database developed using open sources bioinformatics software. Sequence analysis created a huge impact on solanaceae research. using pairwise alignment to find the best matching in query sequences. Fasta format is a text based format for representing either nucleotide sequence or protein sequence (*Higgins, D. G.; Sharp, P. M. (1989)*). The formate originate from the fasta software package. For DNA and Protein it is represented in one letter IUPAC nucleotide code and amino acid code. It is find the local similarity between the sequence and calculates the statistical significance of matches. Mismatch would be connected with a space. Using bioinformatics tools clustalw is a widely used multiple sequence alignment in computer program (*Higgins, D. G.; Bleasby, A. J.; Fuchs, R. (1992)*). An alignment will display by default the following symbols denoting the degree of conservation observed in each column.fasta produce local alignment score the comparison of the query sequence to every sequence in the database. *Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. (1997)*. Sequence alignment or Sequence comparisons lies at the heart of the bioinformatics, which describe the way of arrangement DNA and RNA to identify the regions of similarity among them.

## MATERIALS AND METHODS

The national center of biotechnology information(NCBI) is a multidisciplinary research group that serves as a resources for molecular biology information developing new method to deal with the volume and complexity of data searching and methods that can analyze the structure and function of macromolecules creating computerized systems for storing and analyzing data. The primary database retrieval system at NCBI, which links together several database including gene bank. Fasta is available as a part of a package of program that construct local and global sequence alignment. For a more complete description of fasta and related programs for identifying related DNA/RNA sequence, for evaluating the statistical significance of sequence similarities.

### Database and Corresponding Web services

<u>Database name</u>	<u>Web services type: URL</u>
NCBI	E—Utility web services ( <a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a> )
FASTA	<a href="http://www.ebi.ac.uk/tools">www.ebi.ac.uk/tools</a>
Clustal omega	<a href="http://www.ebi.ac.uk/Tools/msa/clustalw2/">http://www.ebi.ac.uk/Tools/msa/clustalw2/</a>
EMBL/EBI	EMBL-EBI web services ( <a href="http://www.ebi.ac.uk/tools/">http://www.ebi.ac.uk/tools/</a> )
Uniprot KB	Programmatic access services ( <a href="http://www.uniprot.org">http://www.uniprot.org</a> )
EBI/ftp site:	<a href="ftp://ftp.ebi.ac.uk/pub/software/clustalw2/">ftp://ftp.ebi.ac.uk/pub/software/clustalw2/</a>

## RESULTS AND DISCUSSIONS

The FASTA file format now largely used by other sequence database search tools which takes input as nucleotide or protein sequence program (clustalW) clustal is a widely used multiple sequence alignment that manipulate existing alignment, profile analysis and create phylogenetic tree. Alignment can be done by two method slow/accurate, fast/appropriate. Clustal omega is a new multiplae sequence alignment program that high profile technique to generate

alignment between two or more sequences. local sequence alignment program report alignment scores for the alignment constructed, and related(homologous)sequences will have higher alignment scores. The statistical significance of an alignment score is more widely accepted as a metric to comment on the relatedness of the two sequence being aligned. The clustalW and clustalx multiple sequence alignment program have been completely rewritten in C++ (*Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003)*). This facilitate the further development of the alignment algorithms in the future and has proper portion of the program to the latest version of linux, window operating system.(Availability-the program can be run online from the EBI web server. <http://www.ebi.ac.uk/tools/clustalW2>)The clustal series of program are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequence and preparing phylogenetic trees. Clustal was originally developed to run on local computer, numerous web server have been setup, notably at the EBI(European bioinformatics institute).clustalW improving the sensitivity of progressive multiple sequence alignment through sequence weighting ,position specific ,gap penalties. clustalW as a data exploration tools rather than as a definitive analysis method. Multiple sequence alignments are now one of the most widely used bioinformatics analysis. clustalX 2.0 is the new version of the new version of the clustalX graphical alignment tool.

>sequence 1>gi|1015606183|gb|AH001374.2| Solanum lycopersicum chlorophyll a/b-binding protein Cab-3B genes, partial cds

>Sequence2>gi|1015606182|gb|AH001373.2| Solanum lycopersicum chlorophyll a/b-binding protein Cab-3A genes, partial cds

### **Pairwise Statistical Significant Estimation**

Consider the pairwise statistical significance described in obtainable by the following function: where sequence1 and sequence 2, and sc is the scoring scheme(substitution matrix, gap penalties),and N is the number of shuffles.

## **CONCLUSIONS**

ClustalW is very useful in predicting the function and structure of protein/DNA and in identifying new member of protein family. An alignment will display by default the following symbol denoting the degree of conservation observed in each column. Evolutionary relationship can be seen through clado branch or phylobanch.

## **ACKNOWLEDGEMENTS**

We extended our sincere thanks to Dr.Savita “vice chancellor “of Jharkhand rai university, Ranchi, India for kindly providing me the platform to carry out the research.

## **REFERENCES**

1. Altschuhen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David (1990). "Basic local alignment search tool". *Jol, Stepurnal of Molecular Biology* 215 (3): 403–410
2. Andreas D.Baxevanis,B.F.Francis Quellette,"A practical guide to the analysis of Gene and protein .3<sup>RD</sup> Edition october 2004.Published by Wiley,John and Sons.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: Update. *Nucleic Acids Research*, 2004, vol 32, Database Issue: D23-D26. 

4. "ClustalW / ClustalX: Multiple Sequence Alignment". Retrieved 1 October 2013.
5. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003). "Multiple sequence alignment with the Clustal series of programs". Nucleic Acids Res 31 (13): 3497–3500. doi:10.1093/nar/gkg500. PMC 168907. PMID 12824352.
6. Giovannoni, J. 2001, Molecular biology of fruit maturation and ripening, Ann. Rev. Plant Physiol. Plant Mol. Biol., 52, 725-749. [LINK](#)
7. Higgins, D. G.; Sharp, P. M. (1989). "Fast and sensitive multiple sequence alignments on a microcomputer". Computer Applications in the Biosciences (CABIOS) 5 (2): 151–153. doi:10.1093/bioinformatics/5.2.151. PMID 2720464.
8. Higgins, D. G.; Bleasby, A. J.; Fuchs, R. (1992). "CLUSTAL V: Improved software for multiple sequence alignment". Computer Applications in the Biosciences (CABIOS) 8 (2): 189–191. doi:10.1093/bioinformatics/8.2.189. PMID 1591615.
9. Higgins DG, Thompson JD, Gibson TJ. (1996). Using CLUSTAL for multiple sequence alignments. Methods Enzymol., 266, 383-402.
10. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. (1998). Multiple sequence alignment with Clustal X. Trends Biochem Sci., 23, 403-405.
11. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.
12. Madden T. (2002). The NCBI handbook, 2nd edition, Chapter 16, The BLAST Sequence Analysis Tool
13. Mount .David 2004, Bioinformatics:-sequence \$ Genome Analysis", published by Cold spring Harbour laboratory press.
14. NCBI Resource Coordinators (2012). "Database resources of the National Center for Biotechnology Information". Nucleic Acids Research 41 (Database issue): D8–D20.
15. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. (1994). "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". Nucleic Acids Research 22(22): 4673–4680. doi:10.1093/nar/22.22.4673. PMC 308517. PMID 7984417.
16. Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. (1997). "The CLUSTAL\_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools". Nucleic Acids Research 25 (24): 4876–4882. doi:10.1093/nar/25.24.4876. PMC 147148. PMID 9396791.
17. Taylor Willie, Higgins Des 2000, Bioinformatics :Sequence structure and database practical approach “, 1<sup>st</sup> Edition October 2000 , Published by Oxford university press